2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

An Annotation-to-Detection Framework for Autonomous and Robust Vine Trunk Localization in the Field by Mobile Agricultural Robots

Dimitrios Chatziparaschis,[1] Elia Scudiero,[2] Brent Sams,[3] and Konstantinos Karydis[1]

***Abstract*— The dynamic and heterogeneous nature of agricultural fields presents significant challenges for object detection and localization, particularly for autonomous mobile robots that are tasked with surveying previously unseen unstructured environments. Concurrently, there is a growing need for real-time detection systems that do not depend on large-scale manually labeled real-world datasets. In this work, we introduce a comprehensive annotation-to-detection framework designed to train a robust multi-modal detector using limited and partially labeled training data. The proposed methodology incorporates cross-modal annotation transfer and an early-stage sensor fusion pipeline, which, in conjunction with a multi-stage detection architecture, effectively trains and enhances the system's multi-modal detection capabilities. The effectiveness of the framework was demonstrated through vine trunk detection in novel vineyard settings that featured diverse lighting conditions and varying crop densities to validate performance. When integrated with a customized multi-modal LiDAR and Odometry Mapping (LOAM) algorithm and a tree association module, the system demonstrated high-performance trunk localization, successfully identifying over 70% of trees in a single traversal with a mean distance error of less than 0.37 m. The results reveal that by leveraging multi-modal, incremental-stage annotation and training, the proposed framework achieves robust detection performance regardless of limited starting annotations, showcasing its potential for real-world and near-ground agricultural applications.**

I. Introduction

Precision agriculture increasingly relies on systems operating near the ground, including autonomous robots, to improve real-time field monitoring and enable optimized yield prediction and more sustainable operations with reduced labor and costs [1]. A key aspect of autonomous robots performing near-ground proximal sensing tasks is their ability to robustly localize plants based on their distinct characteristics, such as tree trunks, enabling the development of per-plant temporal profiles that are essential for both targeted monitoring and comprehensive field assessments [2]–[5]. However, reliable on-the-go landmark (object of interest) detection remains challenging owing to dynamic environmental conditions, including wind, lighting variability, and interference from dense vegetation [6]. Therefore, developing methods to ensure robust plant (such as trees and grapevines) detection under changing field conditions is essential for fully realizing the benefits of robot-assisted precision agriculture.

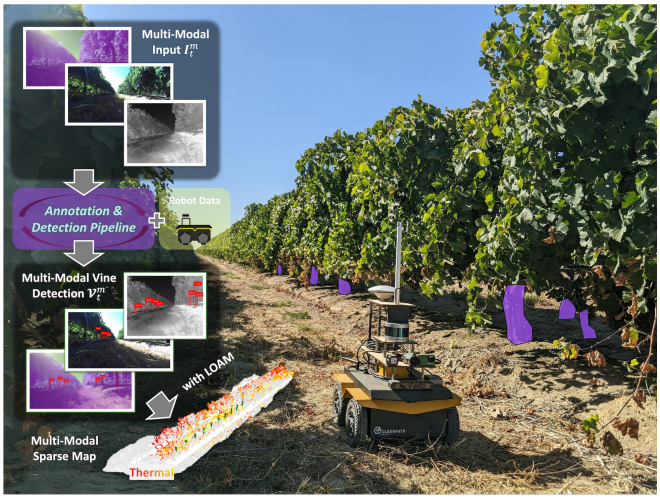


Fig. 1: Our annotation-to-detection framework deployed onboard for multi-modal vine trunk detection. Integration with a customized LOAM framework enables concurrent vine trunk localization and the generation of attribute-rich sparse maps.

Multi-modal sensing has proven beneficial for object detection tasks in agricultural and in-the-wild environments, equipping robots with cross-modal capabilities [7]. The fusion of diverse visual cues that exceed the visible spectrum, along with proximal spatial sensing (such as LiDAR data), can deepen the awareness of robotic operation through attribute-rich mapping, which has been presented in both early and late fusion frameworks [8], [9]. Recent works, such as [10], have incorporated textual and context-aware modalities to further enhance model abilities in object detection and classification. Reliable frameworks for detection and semantic segmentation have used visual modalities with YOLO [11], [12] and Segment Anything Model (SAM) [13], [14] frameworks, and have been applied in the agricultural domain [15], [16]. However, such systems require curated datasets of the given setting for fine-tuning, which are mostly obtained by manual data collection and high-volume human annotation. While these datasets ensure robustness in downstream tasks, their creation remains a bottleneck for field deployment and may limit generalizability across different environments [6], [15].

Semi-Supervised Learning (SSL) has been introduced and applied in cases where large-scale unlabeled data can be eas-

[1] Dept. of Electrical and Computer Engineering, [2] Dept. of Environmental Sciences, Univ. of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA; {dchat013, scudiero, karydis}@ucr.edu, and [3] Gallo, 600 Yosemite Blvd, Modesto, CA 95354, USA; brent.sams@ejgallo.com.

We gratefully acknowledge the support of NSF (#CMMI-2046270, #CNS-2312395, #CMMI-2326309), USDA-NIFA #2024-67022-42532, ONR #N00014-18-1-2252, and the University of California #UC-MRPI M21PR3417. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

ily obtained, via pseudo-labeled data integration [17]. Such approaches have proven effective in the agricultural domain for object detection tasks [18], whereas they leverage multi-modal information to yield performance gains in detection accuracy [19]. Conversely, Few-Shot Learning (FSL) [20], [21] has demonstrated remarkable results when smaller datasets are available, mainly in classification tasks [22], incorporating unseen support datasets by employing similarity criteria during training. Although detection systems have been widely demonstrated in the field, cross-modal approaches must be further examined, particularly when operating in real-time and in-the-wild environments.

In this work, we present a multi-modal annotation-to-detection framework to develop a robust model for cross-modal object detection in the field when limited or no labeled data are available. Our framework integrates a coupled twofold pipeline: 1) utilizing early-stage pseudo-labels from a frozen semantic annotator [14] and association through spatial and visual modalities fusion, and 2) a multi-stage training procedure that uses prior detection knowledge [12] to enrich training datasets and enhance multi-modal detection on-the-go with minimal human intervention. Our cross-modal annotation-to-detection system was deployed on a mobile agricultural robot and evaluated in unseen and vegetation-dense vineyards, specifically for vine trunk detection (Fig. 1). Extensive field evaluation showcased our detector's performance in vine trunk detection, while demonstrating robust tree localization properties when combined with a multi-modal modified version of the AG-LOAM algorithm [23] and an underlying tree trunk association framework [24]. The main contributions of this work are as follows:

- An early-fusion, cross-modal annotation pipeline that generates object-of-interest masks enriched with spatial information and multi-modal attributes (e.g., thermal).
- A stage-incremental detection pipeline utilizing prior knowledge to iteratively refine and include precise pseudo-labels to achieve robust detection performance while minimizing laborious human intervention.
- Accurate in-field vine localization, integrating a multi-modal enhanced LOAM framework [23] with a tree-association algorithm [24], to enable generation of feature-rich sparse point maps accompanied by precise tree detections.

II. Related Works

Multi-modal and data-driven predictive models have shown a positive impact on agricultural applications, spanning from yield prediction to crop detection tasks [15]. Dataset creation and correct annotation are essential for performing any type of data fusion [25], particularly in tasks where more than one sensing modality is employed [26]. Open datasets such as Treescope [27] have been released to advance autonomous mapping and investigate tree phenomics [28], with a particular focus on aerial robotic platforms.

In vineyard settings, the VineSet dataset was published [29] to support multi-modal vine trunk detection, as well as to facilitate grape bunch detection across various growth stages [30]. An initial evaluation was conducted using a mobile robotic platform for vine data acquisition and manual annotation, followed by the deployment of a Tensor Processor Unit (TPU) module for real-time trunk detection [31]. In another study, Magalhães $et\ al.$ [32] employed VineSet to evaluate the real-time performance of various heterogeneous detection platforms. However, while public datasets are vital for model training, operating environments and conditions may differ among robotic applications, leading to poor performance and cross-domain adaptation [33]. Therefore, a more direct, in situ approach is required to leverage the existing knowledge of capable detectors, incorporating multi-modal reasoning while minimizing human intervention.

Vine trunk detection has been a primary and integral component of robotic research, extending it to localization and mapping. Specifically, Papadimitriou $et\ al.$ [34] utilized field uniformity to obtain and spatially characterize valuable visual trunk features to perform loop-closure detection. Slaviček P. $et\ al.$ [35] presented a YOLOv5 framework that was trained and used to obtain enhanced trunk annotations on available open datasets (including VineSet [30]). Specifically, by using a new dataset, a student-teacher network approach for vine trunk detection was employed, and post-filtering with human intervention was applied to exclude faulty detections. Orchard trunk and shrub detection for robotic obstacle avoidance was also demonstrated in [36], with manual collection and labeling of images. Liu Y. $et\ al.$ [37] presented a YOLOv7-based trunk detection for $Camellia\ oleifera$ fruit trees, improving the backbone feature extraction method and training loss for the specific task. Fruit counting on vertical fruiting-wall trees with trunk detection and tracking was also demonstrated by Gao F. $et\ al.$ [38] having initially manual annotation and counting apple fruits and trunks. In all these cases, manual annotations were made for specific tasks, often requiring considerable effort from annotators. Cao Z. $et\ al.$ [39] collected vine trunk datasets and used a Mix-Shelter method for data augmentation with a customized YOLOv8 detector to extract tree lines for safe navigation in orchards. In our work, we present an annotation pipeline as an integral component of our training procedure that aids diverse data augmentation to obtain an accurate in-domain multi-modal detector for vine trunk detection.

III. System Description

A. Sensing Modalities and Specifications

We deployed a Clearpath Jackal mobile robot (Fig. 1) equipped with three fixed-in-place imaging sensors for detection, spanning from visible, Near-Infrared (NIR), to Long-Wave Infrared (LWIR) spectra. The visible modality (RGB) was obtained using a Stereo Labs Zed2i depth camera utilizing the left camera's $720p$ resolution imagery footage. A Mapir Survey3 digital camera was used to capture the NIR data (i.e. through Red-Green-NIR) at a resolution of $720p$, and a FLIR ADK sensor was used to acquire thermal data at 640×512 pixel resolution. Along with the imaging sensors, a Velodyne VLP-16 LiDAR was employed for sparse 3D point

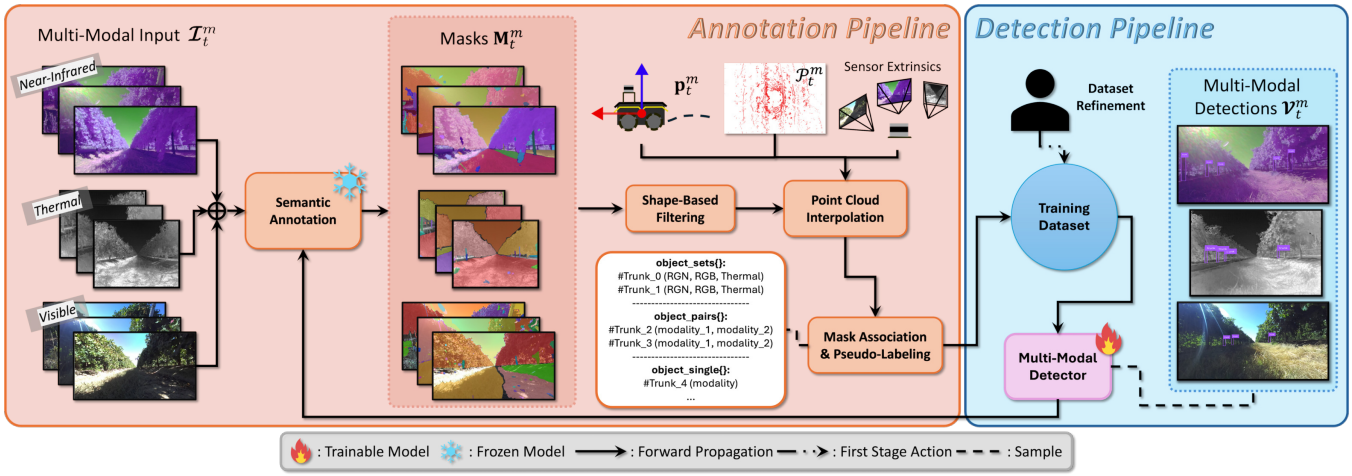


Fig. 2: **Our proposed annotation-to-detection framework for multi-modal vine trunk detection.** The annotation pipeline initiates a tree trunk dataset in the form of annotation masks $\mathbf{M}_t^m$ by combining multi-modal input ($\boldsymbol{\mathcal{I}}_t^v, \boldsymbol{\mathcal{I}}_t^n, \boldsymbol{\mathcal{I}}_t^\tau$) and available spatial information during the robot's operation. The detection pipeline is coupled with the annotated (pseudo-labeled) dataset to iteratively refine and augment the available data, towards training a robust multi-modal vine $\boldsymbol{\mathcal{V}}_t^m$ detector.

cloud generation, which was later used for both mapping and cross-camera detection correspondences. The imaging sensors were set to a 10Hz publishing rate to match the 3D LiDAR operating rate.

The state estimation of the robot in the field was derived from wheel odometry data to determine its pose within a local area. In addition, RTK-GNSS data were used to acquire fixed cm-level global positions using a nearby established RTK-GNSS station. By incorporating the robot's heading, we can apply a forward geodesic transformation to reference each vine detection to the global frame with WGS84 longitude and latitude coordinates. In this way, we generated globally georeferenced AG-LOAM maps from the field, enriched with multi-modal information and the final tree trunk detections.

B. Data Synchronization and Point Cloud Integration

It is important to ensure accurate sensor synchronization and data interpolation to enable reliable detection, particularly when multiple modalities are used. Let $\boldsymbol{\mathcal{I}}_t^m \in \mathbb{R}^{H \times W \times C}$ be the captured image from the m sensing modality with $m \in \{visual, nir, thermal\}$. Each modality might have a different number of C channels, such as the single in thermal image $\boldsymbol{\mathcal{I}}_t^\tau$. Initially, a set of images $\boldsymbol{\mathcal{I}}_t^m$ at time t is obtained using soft synchronization with respect to the common clock of the robot's ROS environment. Together with the set of images ($\boldsymbol{\mathcal{I}}_t^v$, $\boldsymbol{\mathcal{I}}_t^n$, $\boldsymbol{\mathcal{I}}_t^\tau$), the corresponding robot pose $\mathbf{p}_t^m$ is stored as a transformation matrix in $SE(3)$ with respect to the robot's starting position at time t.

Let $\mathcal{P}_{t'} \in \mathbb{R}^3$ be the captured point cloud at time $t' \geq t \in \mathbb{R}^+$. To obtain an accurate point cloud projection for each image $\boldsymbol{\mathcal{I}}_t^m$, we interpolate $\mathcal{P}_{t'}$ to the previous time step t by using the relative transformation from $\mathbf{p}_{t'}$ to $\mathbf{p}_t^m$, owing to robotic movement. Thus, for or each $\boldsymbol{\mathcal{I}}_t^m$ we have,

$$\mathcal{P}_t^m = (\mathbf{p}_t^m)^T \cdot \mathbf{p}_{t'} \cdot \mathcal{P}_{t'} = {}_t^{t'}T \cdot \mathcal{P}_{t'}$$

be the interpolated $\mathcal{P}_{t'}$ from t' to t timestamp, and ${}_t^{t'}T$ is the relative transformation matrix. Since all sensors are affixed to the robot's chassis, their inner transformations are static. We used the KALIBR calibration tool to calculate each sensor's camera matrix K_m and the ACFR library [40] for their extrinsic calibration relative to the VLP-16 LiDAR. Thus, having the transformed point clouds $\mathcal{P}_t^m$, we can use the sensor intrinsics and relative poses to project the points on their image plane of $\boldsymbol{\mathcal{I}}_t^v$, $\boldsymbol{\mathcal{I}}_t^n$, and $\boldsymbol{\mathcal{I}}_t^\tau$, separately.

IV. Proposed Annotation-to-Detection Framework

A. Annotation Pipeline and Pseudo-Labeling

The primary purpose of the annotation pipeline (Fig. 2) is to acquire partial but valuable labels of the object of interest, herein vine trunks, to create a multi-modal training set. We initially used the SAM annotator [14] on the accumulated images from all modalities m to obtain semantic masks $\mathbf{M}_t^m = \{\mathcal{M}_0^m, \mathcal{M}_1^m, ..., \mathcal{M}_i^m\}$ with $\mathcal{M}^m \in \{0,1\}^{H \times W}$. To acquire vine trunk masks over the $\mathbf{M}_t^m$ set, we applied a rectangular shape-based mask filter by extracting the contours of the semantic masks and evaluating whether the number of vertices was greater than two (holding for quadrilateral shapes). By checking the height-to-width contour ratio, we filtered standing rectangle-based masks and considered potential vine trunk masks in the field of view. Figure 3 shows an example of SAM detections on a thermal $\boldsymbol{\mathcal{I}}_t^\tau$ capture. Overlapping masks with an Intersection over Union (IoU) greater than 0.5, or if they occupy more than 40% or less than 5% of the image frame size, are discarded from detection.

As the filtered 2D candidates were acquired from each modality, we proceeded with their spatial association using the corresponding interpolated $\mathcal{P}_t^m$ point clouds. By projecting $\mathcal{P}_t^m$ onto images $\boldsymbol{\mathcal{I}}_t^m$, we form sparse 3D positions for each mask in $\mathbf{M}_t^m$. The centroid points of all lifted masks $\mathbf{M}_t^m$ are computed after we filter out point outliers by

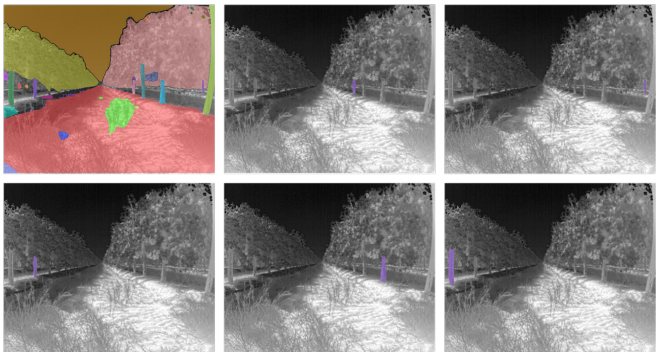


Fig. 3: Generated semantic masks $\mathbf{M}_t^\tau$, based on a thermal image $\boldsymbol{\mathcal{I}}_t^\tau$ input, by using SAM [14] and our shape-based mask filter.

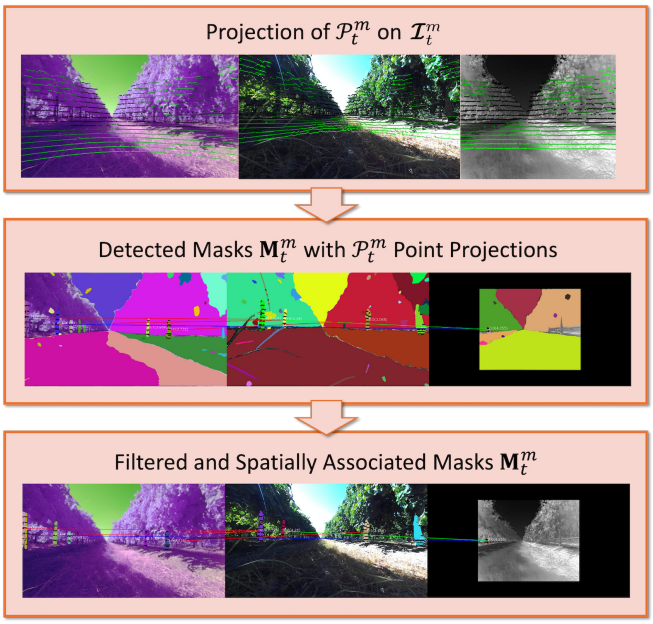


Fig. 4: **Mask association and pseudo-labeling**. By employing the interpolated point clouds $\mathcal{P}_t^m$ and filtered masks $\mathbf{M}_t^m$, the vines are spatially associated to determine if they represent the same object across different modalities.

statistically removing the ones with distance deviation greater than 2σ from the distance mean (i.e. in cases of detection of edge pixels on the mask and/or from LiDAR points that were captured behind the observed object). By iterating over all modalities, we determine the closest cross-modal mask pairs or sets given their 3D centroid positions by having less than 10 cm in 3D distance. We form sets of $\{\mathbf{M}_t^u,\ \mathbf{M}_t^n,\ \mathbf{M}_t^\tau\}$ that correspond to the same detected object across all or partial modalities $\{\boldsymbol{\mathcal{I}}_t^u,\ \boldsymbol{\mathcal{I}}_t^n,\ \boldsymbol{\mathcal{I}}_t^\tau\}$ at time t. Figure 4 shows the intermediate steps for 3D vine association and labeling in our annotation pipeline.

B. Detection Pipeline and Multi-Stage Training

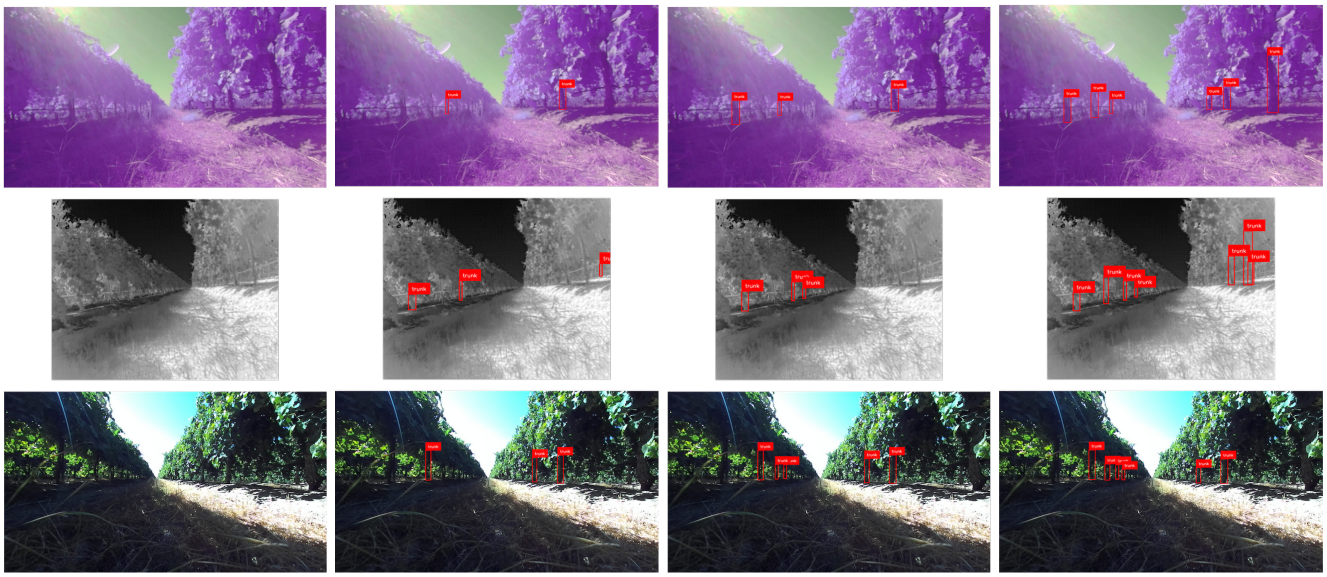
The key goal of the detection pipeline (Fig. 2) is to gradually train a multi-modal detector $\mathcal{D}$ to obtain vine trunks $\boldsymbol{\mathcal{V}}_t^m$ from any given modality, namely $\mathcal{D}(\boldsymbol{\mathcal{I}}_t^m) \rightarrow \boldsymbol{\mathcal{V}}_t^m$ as instance segmentations. Thus, by prioritizing high precision and minimizing false positives, our system iteratively integrates pseudo-labels of detected vines—starting with partially annotated data—during the multi-stage training procedure to build an accurate detection performance. For data collection, we deployed our system in Cabernet Sauvignon vineyards at Gallo Winery ($36°49'49.4''N$, $120°12'36.7''W$).

Initially, our annotation pipeline generated a small tree trunk dataset with 100 image sets of partial vine trunk correspondences using the SAM annotator. The multi-modal dataset was split into training, evaluation, and testing subsets following the 70-20-10 ratio, ensuring an equal ratio among modalities. For the detector, we used the pretrained YOLOv10n [12] model and parsed images from all modalities to enable the development of cross-modal properties. Only in the first step was a human annotator required to remove false annotations in the training set and complete missing data from the evaluation and testing sets. Notably, the manually curated testing set was used to evaluate our detector's performance among the upcoming training stages, alleviating any further human intervention.

TABLE I: Evaluation Stages of Detector $\mathcal{D}$ Aiming Low-False-Positive Vine Trunk Detection

Metrics	$\mathcal{D}_S$	$\mathcal{D}_{S+}$	$\mathcal{D}_T$
Precision	0.02	0.16	**0.83**
Recall	0.14	0.35	**0.53**
mAP_{50}	0.17	0.37	**0.55**
$\text{mAP}_{50:95}$	0.08	0.19	**0.26**

Table I presents the testing results for unseen multi-modal images. As the first-stage detector $\mathcal{D}_S$ was trained on images with limited vine trunk annotations (i.e. filtered SAM masks), lower precision and recall scores were observed. Although the resulting detector, $\mathcal{D}_S$, exhibits limited vine detection capabilities, we leverage its high-confidence predictions in the second iteration to include new trunk annotations within the existing training set. This iterative refinement yielded $\mathcal{D}_{S+}$, with a 60% increase in both recall and precision. In the last step, the refined $\mathcal{D}_{S+}$ model was used to annotate 1500 newly captured multi-modal image sets via the annotation pipeline and repeat training, prioritizing detection precision. The new detector, $\mathcal{D}_T$, has the best performance, scoring 0.83 in precision and 0.53 in recall, as it was trained on the extended training set. Because the detection pipeline prioritizes the reduction of false positives to ensure the veracity of automated pseudo-labeling, it produces fewer but more accurate detections across modalities, with lower mAP scores as a trade-off. Figure 5 illustrates the detection performance of $\mathcal{D}$ models across all training stages on unseen images. We maintained a consistent configuration across all training steps, utilizing a batch size of 32 images for 100 epochs and the pretrained YOLOv10n model as the primary detection backbone. Model evaluation and training were supported by a combination of local computational resources and the National Research Platform (NRP) at UC San Diego.



(a) Raw $\mathcal{I}_t^m$ Inputs (b) $\mathcal{V}_t^m$ with $\mathcal{D}_S$ (c) $\mathcal{V}_t^m$ with $\mathcal{D}_{S+}$ (d) $\mathcal{V}_t^m$ with $\mathcal{D}_T$

Fig. 5: **Evolution of vine trunk detection.** By transitioning from the $\mathcal{D}_S$ model, trained on the partially annotated dataset (b), to $\mathcal{D}_T$ model, which incorporates the enriched pseudo-labeled dataset (d), the detector achieves superior performance across all input modalities.

TABLE II: Evaluation of Vine Trunk Localization

Experiment	Single-Row with 10 Trees			Dual-Row with 7×7 Trees
Metrics	$1st - 5th$ Tree	$5th - 10th$ Tree	Total	Total
L^2 Distance Error	0.32 m	0.24 m	0.28 m	0.37 m
$\text{MAE}_{<0.5}$	0.05 m	0.11 m	0.09 m	0.06 m
$\text{RMSE}_{<0.5}$	0.05 m	0.12 m	0.10 m	0.07 m
$\text{Recall}_{<0.5}$	60%	80%	80%	71%
Total Tree Detections	4 out of 5	4 out of 5	8 out of 10	10 out of 14

V. Experimental Evaluation

To evaluate vine trunk detection and localization, we deployed our system across two unseen fields at Gallo Vineyards and conducted surveys between straight-line tree rows. By maintaining westward heading during post-noon hours, we examined our system's ability in vine trunk detection and localization across different numbers of trees and sunlight conditions within the rows. To derive a sparse 3D map of the surveyed vineyard, we modified the AG-LOAM [23] to register information from the additional sensing modalities during LiDAR mapping. Along with the mapping system, we integrated a multi-modal tree localization module [24] for on-the-go and by-tree information tracking using our online trunk detections as an input. The positions of the detected trees were evaluated against the surveyed ground truth. The localization results are presented in Table II.

A. Single-Row Vine Detection

In this experimental setup, we evaluated the system on a 20×5 m^2 vineyard row containing 10 vine trees located to the right of the robot's trajectory. As a single tree row, all vines were uniformly exposed to sunlight, yielding a consistent lighting scenario. During the survey, the robot maintained a constant linear speed of 0.5 m/s and captured multi-modal data along the path, containing vines at varying distances, angles, and partially occluded instances.

Initially, as shown in Table II, our system was able to correctly detect vine trunks, identifying more than 80% of the trees with a single pass from the robot. Specifically, in both the first and second sets of five trees, our detector demonstrated consistency scoring 0.32 m and 0.24 m of Euclidean (L^2) distance error with respect to the ground truth, respectively. Figure 6 illustrates tree trunk detections that are notably close to the ground truth tree positions. False-positive detections occurred when the trunk points fell in front of or behind the actual trunk points because of the deep vegetation around the observed vine, thus forming small clusters of false-positive points. The high recall rates across both five tree groups and the total 10 tree assessments demonstrate our system's reliability, exceeding 80% in recall and minimizing the occurrence of false-negative trunk masks/detections. Notably, successful vine trunk detections were considered if they fell within a 3D Euclidean distance threshold of 0.15 m from the ground truth trunk positions.

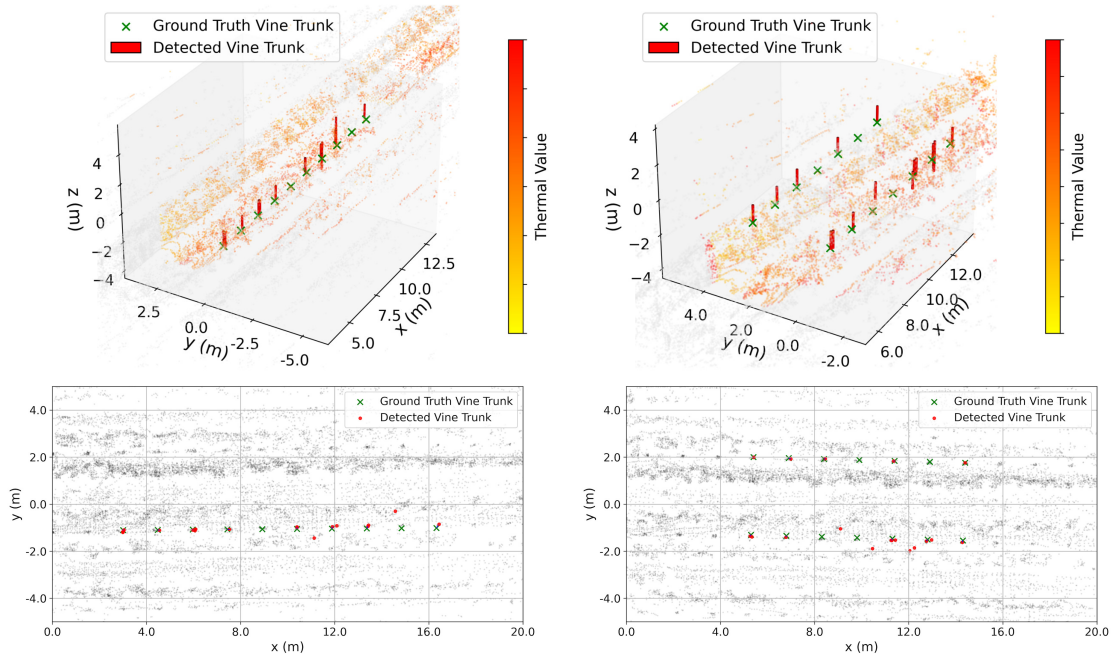


Fig. 6: **Field assessment of vine detection and localization.** Our system accurately detects and localizes tree trunks in a single pass across both single-row (10 trees, left) and dual-row (7×7 trees, right) configurations. By incorporating multi-modal information into the LOAM registration process, the system generates feature-rich sparse maps—such as the illustrated thermal visualizations—simultaneously with real-time vine trunk detection.

B. Dual-Row Vine Detection

Following the evaluation setup of the single-vine row, we accessed our system on a 15×5 m^2 vineyard, where we enabled vine detections from the sides of the robot. In this way, the vine detector considers all appearing trees in the sensors' field of view during traverse. Herein, the detection task is more susceptible owing to the varying shadowing effects arising from the raycasted sunlight within vine rows (seen also in Figures 5) during the post-noon survey hours.

As demonstrated by the results in Table II, our detector effectively acquires trunk detections by obtaining 10 out of 14 trees with a single field pass. Figure 6 also illustrates tree trunk detection in the local LOAM frame, along with the ground truth positions. Overall, our system achieves less than 0.37 m of L^2 error for each inspected tree by having less than 0.07 m of detection error measured in the Root Mean Squared Error (RMSE) in the candidate detections of 0.50 m radius surrounding the target trees. Through a single-pass field traversal and multi-modal detection of each trunk, our system demonstrates a consistently high recall rate—obtaining more than 70% trees—even in the more variable sunlight assessment.

VI. Conclusion

This paper presents an onboard annotation-to-detection framework for identifying multi-modal objects of interest in the field, herein vine trunks, in scenarios where labeled data are limited. Our annotation pipeline leverages reliable and refined detectors to generate object pseudo-labels, which associate spatially across different modalities, to enrich a multi-modal training dataset. Simultaneously, the detection pipeline employs a multi-stage training procedure to iteratively refine the detector across different modalities, requiring minimal human intervention only in the early stage. Our system demonstrated robust vine trunk detection capabilities even when starting with a small curated dataset, obtaining more than 70% of the trees in different vineyard settings with a single pass from the robot. Integration with LOAM and a tree trunk localization framework yielded a distance localization error below 0.37 m, enabling the generation of multi-modal sparse maps. The future implications of this work include scalable, multi-agent object detection that fuses cross-modal and cross-agent information to maximize detection accuracy and alleviate the need for human supervision during labor-intensive labeling tasks in the field.

References

[1] R. Sparrow and M. Howard, "Robots in agriculture: prospects, impacts, ethics, and policy," *precision agriculture*, vol. 22, pp. 818–833, 2021.

[2] F. Morbidini, A. Samanta, C. Maucieri, K. Karydis, P. A. Mauk, T. H. Skaggs, and E. Scudiero, "Robotic mapping of soil volumetric water content with geospatial soil apparent electrical conductivity in micro-irrigated citrus orchards in california," *Computers and Electronics in Agriculture*, vol. 245, p. 111540, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169926001353

[3] E. Hyyppä, J. Hyyppä, T. Hakala, A. Kukko, M. A. Wulder, J. C. White, J. Pyörälä, X. Yu, Y. Wang, J.-P. Virtanen *et al.*, "Under-canopy uav laser scanning for accurate forest field measurements," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 164, pp. 41–60, 2020.

[4] N. Niknejad, R. Bidese-Puhl, Y. Bao, K. G. Payn, and J. Zheng, "Phenotyping of architecture traits of loblolly pine trees using stereo machine vision and deep learning: Stem diameter, branch angle, and

branch diameter," *Computers and Electronics in Agriculture*, vol. 211, p. 107999, 2023.

[5] E. Scudiero, A. Singh, G. R. Mahajan, D. Chatziparaschis, J. Banik, K. Karydis, D. A. Houtz, and T. H. Skaggs, "Near-ground microwave radiometry for on-the-go surface soil moisture sensing in micro-irrigated orchards in california," *Agrosystems, Geosciences & Environment*, vol. 8, no. 3, p. e70202, 2025. [Online]. Available: https://acsess.onlinelibrary.wiley.com/doi/abs/10.1002/agg2.70202

[6] O. Wosner, G. Farjon, and A. Bar-Hillel, "Object detection in agricultural contexts: A multiple resolution benchmark and comparison to human," *Computers and Electronics in Agriculture*, vol. 189, p. 106404, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016816992100421X

[7] C. S. Parr, D. G. Lemay, C. L. Owen, M. J. Woodward-Greene, and J. Sun, "Multimodal ai to improve agriculture," *IT Professional*, vol. 23, no. 3, pp. 53–57, 2021.

[8] S. A. Zamani and Y. Baleghi, "Early/late fusion structures with optimized feature selection for weed detection using visible and thermal images of paddy fields," *Precision Agriculture*, vol. 24, no. 2, pp. 482–510, 2023.

[9] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 2020, pp. 1–6.

[10] Y. Lu, X. Lu, L. Zheng, M. Sun, S. Chen, B. Chen, T. Wang, J. Yang, and C. Lv, "Application of multimodal transformer model in intelligent agricultural disease detection and question-answering systems," *Plants*, vol. 13, no. 7, p. 972, 2024.

[11] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[12] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," 2024. [Online]. Available: https://arxiv.org/abs/2405.14458

[13] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[15] M. El Sakka, M. Ivanovici, L. Chaari, and J. Mothe, "A review of cnn applications in smart agriculture using multimodal data," *Sensors*, vol. 25, no. 2, p. 472, 2025.

[16] C. M. Badgujar, A. Poulose, and H. Gan, "Agricultural object detection with you only look once (yolo) algorithm: A bibliometric and systematic literature review," *Computers and Electronics in Agriculture*, vol. 223, p. 109090, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169924004812

[17] G. Huang, I. Laradji, D. Vazquez, S. Lacoste-Julien, and P. Rodriguez, "A survey of self-supervised and few-shot object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4071–4089, 2022.

[18] G. Tseng, K. Sinkovics, T. Watsham, D. Rolnick, and T. C. Walters, "Semi-supervised object detection for agriculture," in *2nd AAAI Workshop on AI for Agriculture and Food Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=AR4SAOzcuz

[19] C. Devaguptapu, N. Akolekar, M. M Sharma, and V. N Balasubramanian, "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[20] Z. Xin, S. Chen, T. Wu, Y. Shao, W. Ding, and X. You, "Few-shot object detection: Research advances and challenges," *Information Fusion*, vol. 107, p. 102307, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156625352400085X

[21] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, Jun. 2020. [Online]. Available: https://doi.org/10.1145/3386252

[22] N. Ragu and J. Teo, "Object detection and classification using few-shot learning in smart agriculture: A scoping mini review," *Frontiers in Sustainable Food Systems*, vol. Volume 6 - 2022, 2023. [Online]. Available: https://www.frontiersin.org/journals/sustainable-food-systems/articles/10.3389/fsufs.2022.1039299

[23] H. Teng, Y. Wang, D. Chatziparaschis, and K. Karydis, "Adaptive lidar odometry and mapping for autonomous agricultural mobile robots in unmanned farms," *Computers and Electronics in Agriculture*, vol. 232, p. 110023, 2025.

[24] D. Chatziparaschis, H. Teng, Y. Wang, P. Peiris, E. Seudiero, and K. Karydis, "On-the-go tree detection and geometric traits estimation with ground mobile robots in fruit tree groves," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 15 840–15 846.

[25] H. B. Mitchell, *Image fusion: theories, techniques and applications*. Springer Science & Business Media, 2010.

[26] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2021.

[27] D. Cheng, F. Cladera, A. Prabhu, X. Liu, A. Zhu, P. C. Green, R. Ehsani, P. Chaudhari, and V. Kumar, "Treescope: An agricultural robotics dataset for lidar-based mapping of trees in forests and orchards," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14 860–14 866.

[28] D. Houle, D. R. Govindaraju, and S. Omholt, "Phenomics: the next challenge," *Nature reviews genetics*, vol. 11, no. 12, pp. 855–866, 2010.

[29] A. S. Aguiar, N. N. Monteiro, F. N. d. Santos, E. J. Solteiro Pires, D. Silva, A. J. Sousa, and J. Boaventura-Cunha, "Bringing semantics to the vineyard: An approach on deep learning-based vine trunk detection," *Agriculture*, vol. 11, no. 2, p. 131, 2021.

[30] A. S. Aguiar, S. A. Magalhães, F. N. Dos Santos, L. Castro, T. Pinho, J. Valente, R. Martins, and J. Boaventura-Cunha, "Grape bunch detection at different growth stages using deep learning quantized models," *Agronomy*, vol. 11, no. 9, p. 1890, 2021.

[31] A. S. Aguiar, F. N. Dos Santos, A. J. M. De Sousa, P. M. Oliveira, and L. C. Santos, "Visual trunk detection using transfer learning and a deep learning-based coprocessor," *IEEE Access*, vol. 8, pp. 77 308–77 320, 2020.

[32] S. C. Magalhães, F. N. dos Santos, P. Machado, A. P. Moreira, and J. Dias, "Benchmarking edge computing devices for grape bunches and trunks detection using accelerated object detection single shot multibox deep learning models," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105604, 2023.

[33] P. Oza, V. A. Sindagi, V. VS, and V. M. Patel, "Unsupervised domain adaptation of object detectors: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4018–4040, 2024.

[34] A. Papadimitriou, I. Kleitsiotis, I. Kostavelis, I. Mariolis, D. Giakoumis, S. Likothanassis, and D. Tzovaras, "Loop closure detection and slam in vineyards with deep semantic cues," in *International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2251–2258.

[35] P. Slaviček, I. Hrabar, and Z. Kovačić, "Generating a dataset for semantic segmentation of vine trunks in vineyards using semi-supervised learning and object detection," *Robotics*, vol. 13, no. 2, p. 20, 2024.

[36] L. Xiangyang, H. Lei, G. Shiyong, and Z. Liuqun, "Analysis of construction orchard trunk shrub dataset and target detection," in *International Conference on Advanced Mechatronic Systems (ICAMechS)*, 2024, pp. 326–330.

[37] Y. Liu, H. Wang, Y. Liu, Y. Luo, H. Li, H. Chen, K. Liao, and L. Li, "A trunk detection method for camellia oleifera fruit harvesting robot based on improved yolov7," *Forests*, vol. 14, no. 7, p. 1453, 2023.

[38] F. Gao, W. Fang, X. Sun, Z. Wu, G. Zhao, G. Li, R. Li, L. Fu, and Q. Zhang, "A novel apple fruit detection and counting methodology based on deep learning and trunk tracking in modern orchard," *Computers and Electronics in Agriculture*, vol. 197, p. 107000, 2022.

[39] Z. Cao, C. Gong, J. Meng, Y. Rao, W. Hou *et al.*, "Orchard vision navigation line extraction based on yolov8-trunk detection," *IEEE Access*, 2024.

[40] D. Tsai, S. Worrall, M. Shan, A. Lohr, and E. Nebot, "Optimising the selection of samples for robust lidar camera calibration," in *IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 2631–2638.